Web Archiving in the United States: A 2017 Survey

An NDSA Report

NDSA

Results of a Survey of Organizations Preserving Web Content

October 2018

AUTHORS
Matthew Farrell, Duke University
Edward McCain, University of Missouri
Maria Praetzellis, Internet Archive
Grace Thomas, Library of Congress
Paige Walker, Boston College

TABLE OF CONTENTS

ABOUT THE NATIONAL DIGITAL STEWARDSHIP ALLIANCE

Founded in 2010, the National Digital Stewardship Alliance (NDSA) is a consortium of institutions that are committed to the long-term preservation of digital information. NDSA's mission is to establish, maintain, and advance the capacity to preserve our nation's digital resources for the benefit of present and future generations. NDSA member institutions represent all sectors, and include universities, consortia, non-profits, professional associations, commercial enterprises, and government agencies at the federal, state, and local levels.

More information about the NDSA is available at http://www.ndsa.org.

DOI: 10.17605/OSF.IO/3QH6N

INTRODUCTION

From October 2 to November 20, 2017, a working group of individuals representing multiple NDSA member institutions and interest groups conducted a survey of organizations in the United States actively involved in, or planning to start, programs to archive content from the Web. This effort builds upon and extends a broader effort begun in three earlier surveys,[1] which the NDSA Web Archiving Survey working group has conducted since 2011.[2, 3, 4] The goal of these surveys is to better understand the landscape of Web archiving activities in the United States by investigating the organizations involved; the history and scope of their Web archiving programs; the types of Web content being preserved; the tools and services being used; access and discovery services being offered; and overall policies related to Web archiving programs. The responses from this survey document the current state of U.S. Web archiving initiatives and the comparison with the results of the 2011, 2013, and 2016 surveys enables an analysis of emerging trends. This report describes the current state of the field, tracks the evolution of the field over the last few years, and points to future opportunities and developments.

METHODOLOGY

Volunteers from the NDSA community formed the 2017 Web Archiving Survey working group in March 2017. The working group self-organized and began by reviewing previous surveys in order to identify any needed changes or additions to questions. The group was careful to make sure historical comparisons could be made from the 2011, 2013, and 2016 surveys in order to ensure the integrity of longitudinal data. A few questions and response options were added to improve clarity and to address new issues or practices. For example, two questions relating to social media capture were added in the Tools section. In the case of question 23, which asks “If you are (or have been) using an external service for data capture, which one(s) do you use?” five of the choices for services were eliminated because no responses for those services were collected during the 2016 survey. Responses to one question in the Access and Discovery section were excluded from this report due to concerns about a lack of clarity in the wording of the survey.[5]

[1] The first survey was held in late 2011 with results published online in June 2012, the second held in late 2013 with results published online in September 2014, and the most recent survey conducted in 2016 with results published online in February 2017.

[2] “Web Archiving Survey Report,” *NDSA Report*, June 19, 2012, accessed July 6, 2018, http://www.digitalpreservation.gov/documents/ndsa_web_archiving_survey_report_2012.pdf.

[3] “Web Archiving in the United States: A 2013 Survey,” *NDSA Report*, September 2014, accessed July 6, 2018, http://ndsa.org/documents/NDSA_USWebArchivingSurvey_2013.pdf.

[4] “Web Archiving in the United States: A 2016 Survey,” *NDSA Report*, February 2017, accessed July 6, 2018, https://ndsa.org/documents/WebArchivingintheUnitedStates_A2016Survey.pdf.

[5] Responses from question 27: “What kind(s) of access does your organization itself provide? Check all that apply.” in the Access and Discovery section were excluded from this report. After much deliberation from the survey working group and reviewers, it was decided that the wording of the question was misleading and highly likely that respondents interpreted the question differently. This section will be under high priority to restructure in future reports as access and discovery are crucial aspects of the web archiving workflow.

The survey was available from October 2 to November 20, 2017 via SurveyMonkey. Survey respondents were solicited through mailing lists, blogs, social media and other channels. After the survey closed, the working group reviewed a total of 156 responses and removed 39 responses that were tests, substantially incomplete, or from outside the United States, leaving a total of 119 responses for analysis. Respondents were not required to answer all questions. Percentages reported for individual questions reflect the number of responses to that question rather than the total number of respondents participating in the survey.

About the Survey

The 2017 NDSA Web Archiving Survey asked 39 questions organized around five distinct topic areas: background information about the respondent's organization; details about the current state of their Web archiving program; tools and services used by their program; access and discovery systems and approaches; and program policies involving capture, availability, and types of Web content.

The working group shared the survey on a number of listservs.[6] The working group also promoted the survey on Archive-It's blog and the Library of Congress's blog, *The Signal*. On Twitter, the working group promoted the survey from their personal accounts, the Archive-It account, and the NDSA account.

This report analyzes 119 responses (n=119), excluding responses that were not completed or completed by organizations outside the United States. This represents an increase of 14% over the 2016 survey's 104 responses, an increase of 29% over the 2013 survey's 92 responses, and an increase of 55% over the 2011 survey's 77 responses. The survey consisted primarily of multiple choice questions, with some questions also containing free text response fields for clarification or elaboration of answers.[7]

RESPONDENT CHARACTERISTICS

The first category of questions asked respondents to characterize their organization's type, to highlight any ties to Web archiving community groups, and to indicate the status of their Web archiving activity.

[6] NDSA-ALL LISTSERV; DLF-ANNOUNCE LISTSERV; Code4Lib LISTSERV; Research Data Access and Preservation LISTSERV; Boston Digital Humanities LISTSERV; Archive-It LISTSERV; Society of American Archivists Web Archiving Section, Electronic Records Section, and general announcement LISTSERVs; Association of College & Research Libraries Digital Humanities LISTSERV; Association of Southeastern Research Libraries LISTSERV; the BitCurator Consortium LISTSERV; and the Digital Preservation Library of Congress LISTSERV.

[7] Survey instrument and anonymized survey data are available via the Internet Archive at https://archive.org/details/2017ndsasurvey_201809.

Organization Type

The types of organizations represented in the responses are consistent with the three previous surveys, with one notable increase. Organizations identifying as public libraries increased to 13% (15 of 119) of respondents from less than 3% (2, 2, and 1 organizations in 2011, 2013, and 2016, respectively) of respondents in each of the previous surveys. The increase in public libraries can be traced to the Internet Archive's Community Webs program, funded by the Institute of Museum and Library Services, the Kahle-Austin Foundation, and Archive-It, that kicked off in 2017 and is providing training and technical services to public libraries in the field of Web archiving for local history preservation.[8] Respondents identifying as a College or University remain the majority of respondents at 61% (72 of 119).

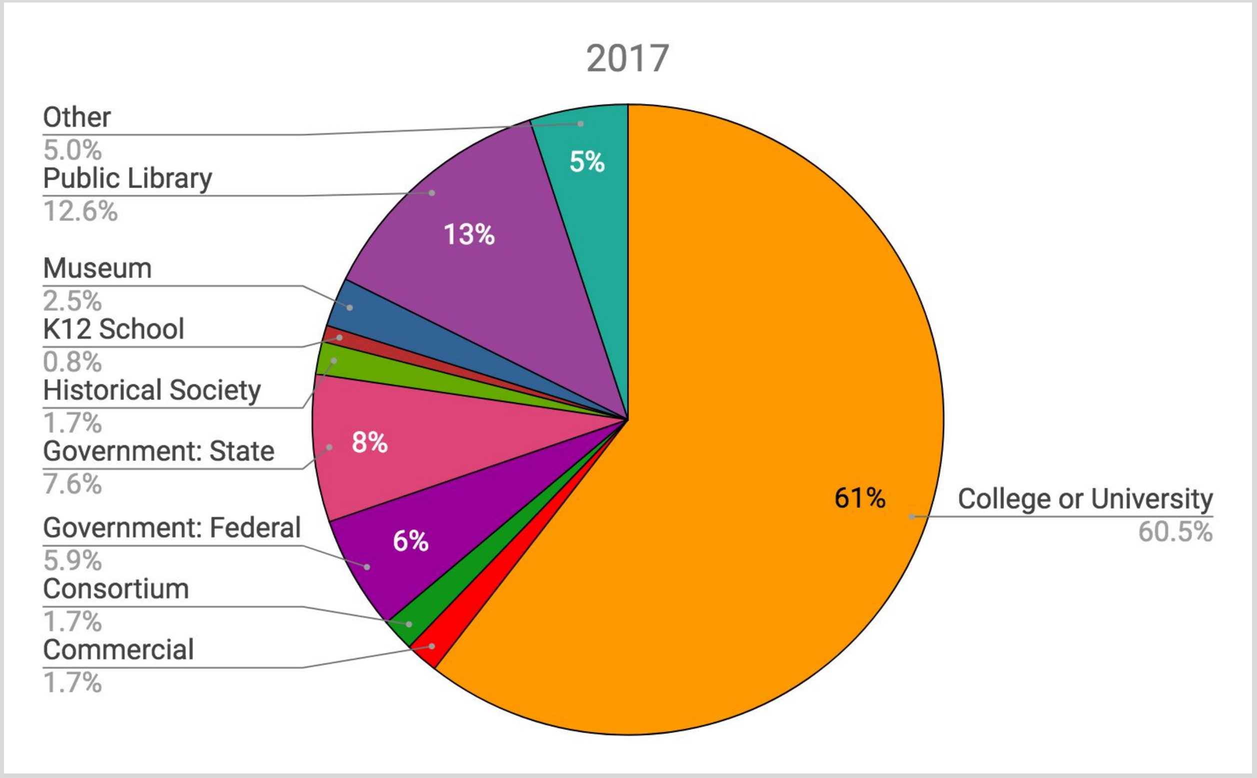

FIGURE 1: Responding organization type in 2017

[8] Jefferson Bailey, "IMLS Grant to Advance Web Archiving in Public Libraries," Internet Archive Blogs, July 18, 2017, accessed August 22, 2018, https://blog.archive.org/2017/07/18/imls-grant-to-advance-Web-archiving-in-public-libraries/.

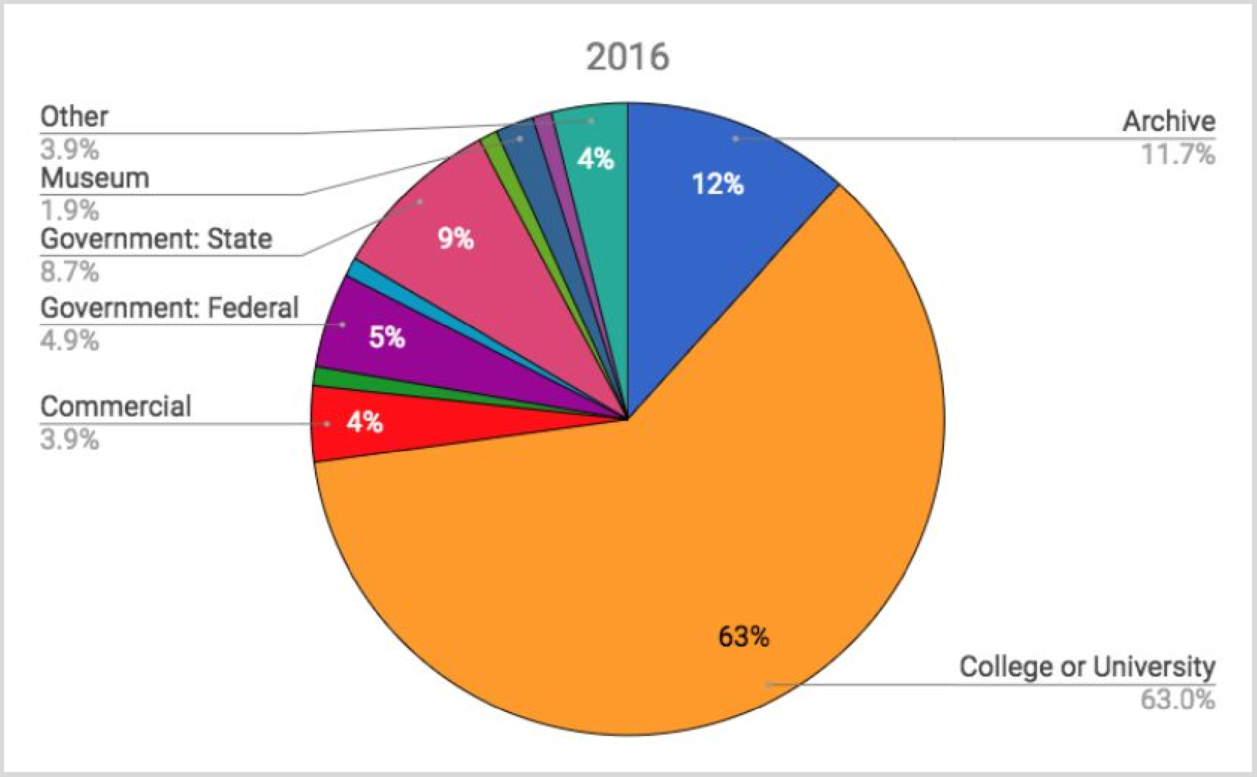

FIGURE 2: Responding organization type in 2016

Group Affiliations

The United States Web archiving community is highly engaged in Web archiving-related professional groups. Over 71% (59 of 82) of survey respondents reported affiliation with the Society of American Archivists' Web Archiving Section (SAA WebArchRT), consistently making it the group with highest affiliation over all survey periods. 2017 is the first year this survey has offered Digital Library Federation (DLF) membership as an option for respondents, which resulted in over half, 54% (44 of 82), of respondents reporting affiliation. Reported NDSA membership fell slightly from 2016, making up only 45% (37 of 82) of respondents, compared to 53% (41 of 77) of respondents in 2016. Affiliation with the International Internet Preservation Consortium (IIPC) remains the professional group with least affiliation among the United States Web archiving community, with 10% (8 of 82) of respondents reporting membership. These metrics are likely due to the varying costs and missions of these diverse organizations. Membership among multiple groups continued to increase in 2017. Organizations reporting affiliation with two or more professional associations jumped from 30% in 2016 (23 of 77) to 43% (35 of 82) of respondents in 2017.

Activity Status

The survey asked respondents to indicate the overall activity status of their Web archiving program, choosing among four stages: Planning (considering archiving but haven't started yet), Pilot (testing), Production (actively capturing), and Discontinued (have collected content in the past but are not currently collecting). As shown in Figure 3, the percentage of respondents reporting their program in Production status has continued the steady increase shown over the previous surveys. Currently, 81% (96 of 119) of respondents consider themselves in Production status.

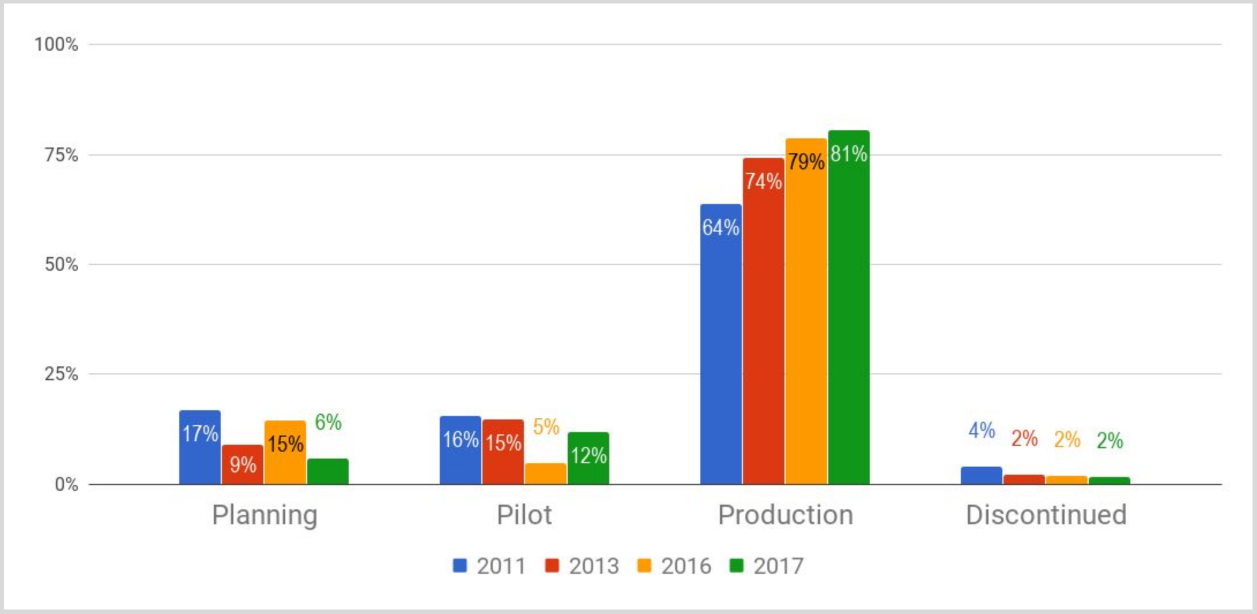

FIGURE 3: Status of Web archiving activity 2011-2017

Throughout the previous surveys, respondents reporting their program in Planning status has changed each year. 2017 was a down year for respondents in Planning status representing 6% (7 of 119) of respondents; however, respondents reporting their program in Pilot status (12%, or 14 of 119) rose since 2016 (5%, or 5 of 103). Since the number of programs in Planning status was up to 15% (15 of 103) in 2016, this could indicate that some respondents in Planning stages during 2016 had progressed to Pilot or Production status in 2017. These metrics could also indicate that more programs move directly to Pilot or Production stages, rather than Planning.

PROGRAM INFORMATION

This section of the report addresses questions related to the Web archiving program of the respondents, including goals, perceptions of progress, staffing levels and skills, content areas of concern, and collaboration.

When Programs Started

The growing response rate indicates that Web archiving remains a growth area for many organizations. Twenty-seven percent (30 of 110) of survey respondents began their Web archiving program in the time since the last survey was completed in 2016. This figure is consistent with earlier results. In the three previous surveys, programs beginning recently (the most immediate preceding two year period) accounted for a range from 29% (24 of 82) to 38% (31 of 81).

Goals and Development

The responses to this question are consistent with surveys from 2013 and 2016, with 86% (102 of 119) of respondents archiving their own institution's Web content, 56% (67 of 119) archiving third-party content, and 41% (49 of 119) capturing both their own content as well as third-party content. While respondents reported archiving third-party content in 2016 and 2017 at nearly equal rates (56% or 50 of 89, and 56% or 67 of 119, respectively), this figure is significantly lower than the 2011 rate (80%, 57 of 81).

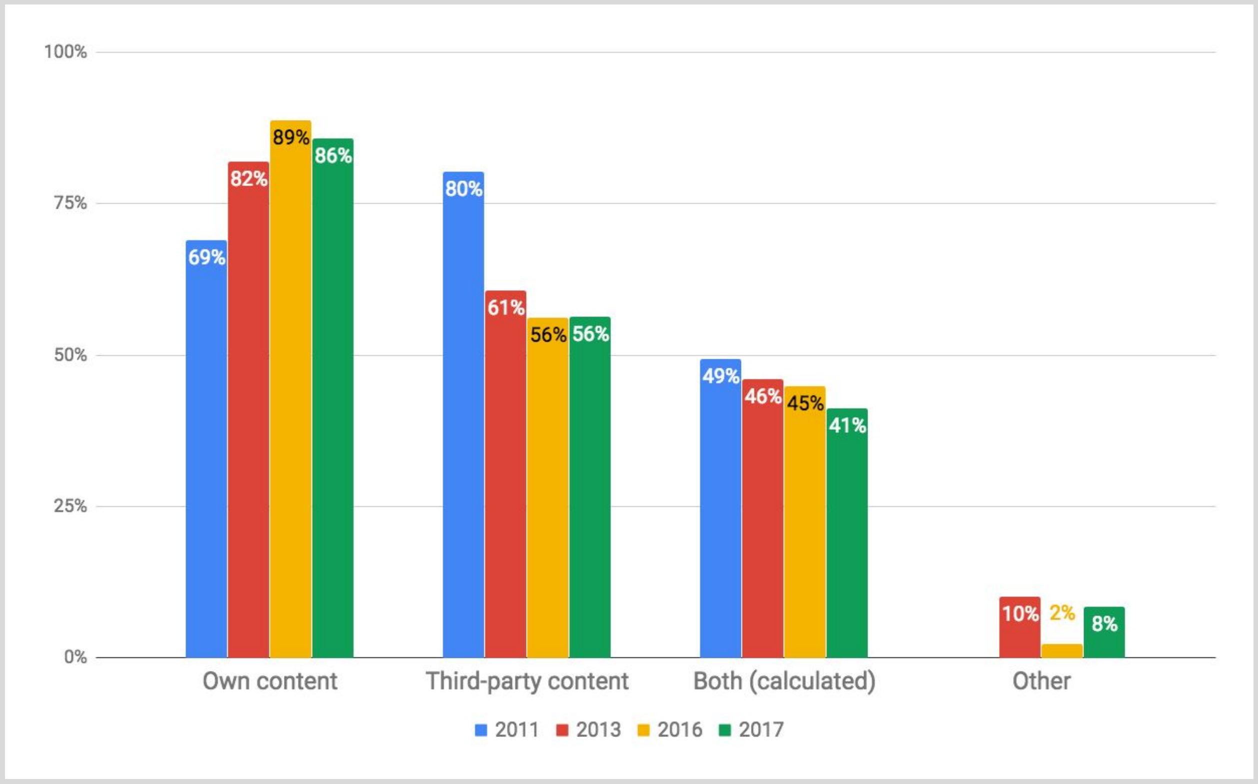

FIGURE 4: Institutional goals for collecting content 2011-2017

This question also had an option for "Other," which presented a free text response. Ten respondents provided additional information in this option. Some provided similar responses, with three distinct respondents mentioned archiving websites related to their local community, while two respondents mentioned Web archiving related to events.

Perceptions of Progress

Nearly 80% of survey respondents reported making progress in their Web archiving programs over the past two years: 40% (48 of 119) reported significant progress; 38% (45 of 119) reported some progress. These are similar to the results of the 2016 and 2013 survey, when the question was first asked. More interesting might be the relatively flat rate of institutions responding that they were at about the same status as two years ago: 19% (18 of 93) in 2013, 20% (18 of 88) in 2016, and 20% (24 of 119) in 2017. The optimistic interpretation might be that a number of institutions feel their program has reached maturity and remained there, but there may be other possibilities, such as that barriers that existed two years ago may still exist either for individual institutions or across the Web archiving field.

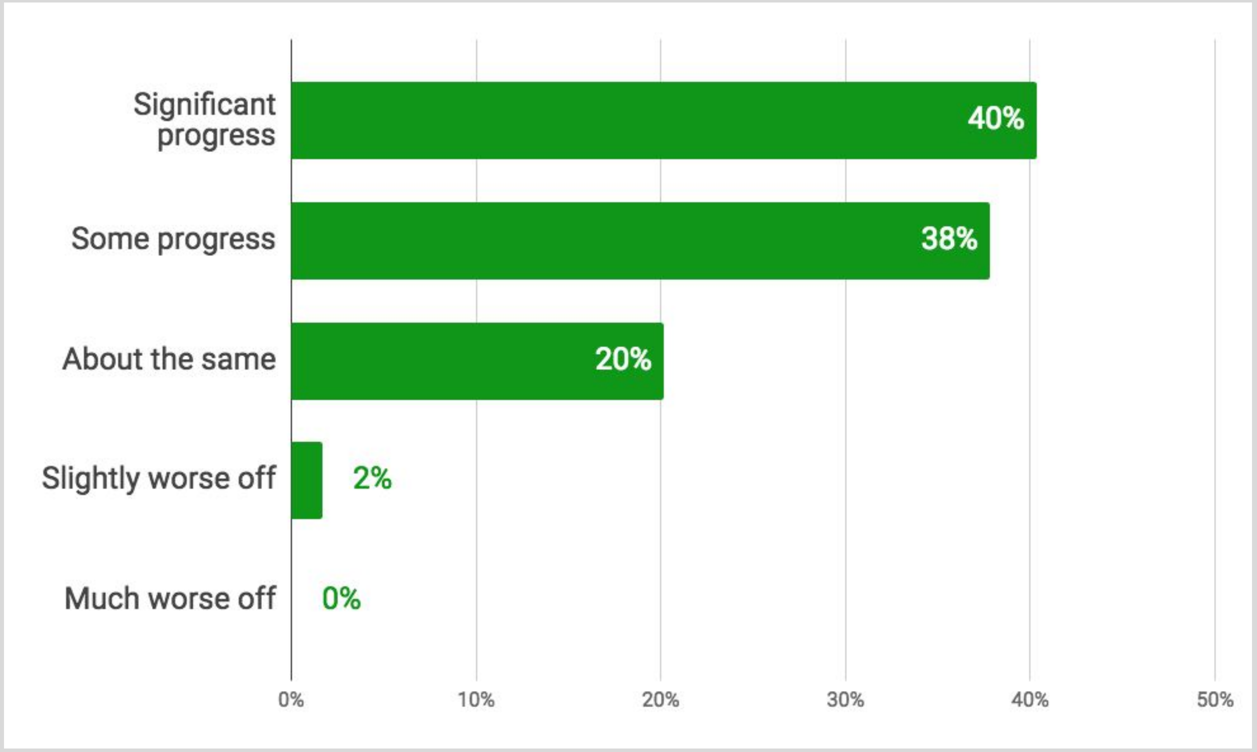

FIGURE 5: Perceptions of overall progress in the last two years

An optional follow-up offered a free text field to “provide an alternate answer or commentary on [the] answer to [the previous question].” Thirty-three respondents took the opportunity to offer additional, anecdotal information. Three institutions described hiring staff members to start or assist in Web archiving. Eight institutions described launching a new program in the last two years, while seven institutions described expanding their program or the scope of their program. Three institutions described a program that was either treading water or had lost momentum due to staff departures.

Another question on the survey asked respondents to consider their progress in the dimensions of the Web Archiving Lifecycle Model,[9] choosing the top three areas where their

[9] The Web Archiving Lifecycle Model was developed by the Archive-It team at the Internet Archive in 2013. The model was developed in order to address the lack of best practices and to increase awareness of the importance of Web archiving as fundamental to digital preservation. The Archive-It

programs have made the most and least progress. With respect to making the most progress, this year most respondents ranked Data Capture (54%, 64 of 119), Appraisal & Selection (47%, 56 of 119), and Scoping (34%, 40 of 119) highest. Compared to the 2016 survey, Data Capture and Appraisal & Selection ranked at the top, but respondents selected Vision & Objectives in higher numbers in 2016 (40%, 34 of 85) than in 2017 (24%, 29 of 119). The other dimension seeing a notable difference from 2016 to 2017 is Resources & Workflow, which saw a smaller rate of selection in 2017 (24%, 29 of 119) than 2016 (34%, 29 of 85). Vision & Objectives and Resources & Workflow are two dimensions associated with policy development, and potentially with the planning and launch stages of a Web archiving program. The decline in selection rates for these two dimensions may indicate that many Web archiving programs are reaching maturity. On the other hand, the activities associated with appraisal and selection, data capture, and scoping are often skills that develop with experience, so there is some logic to their presence with the highest response rate.

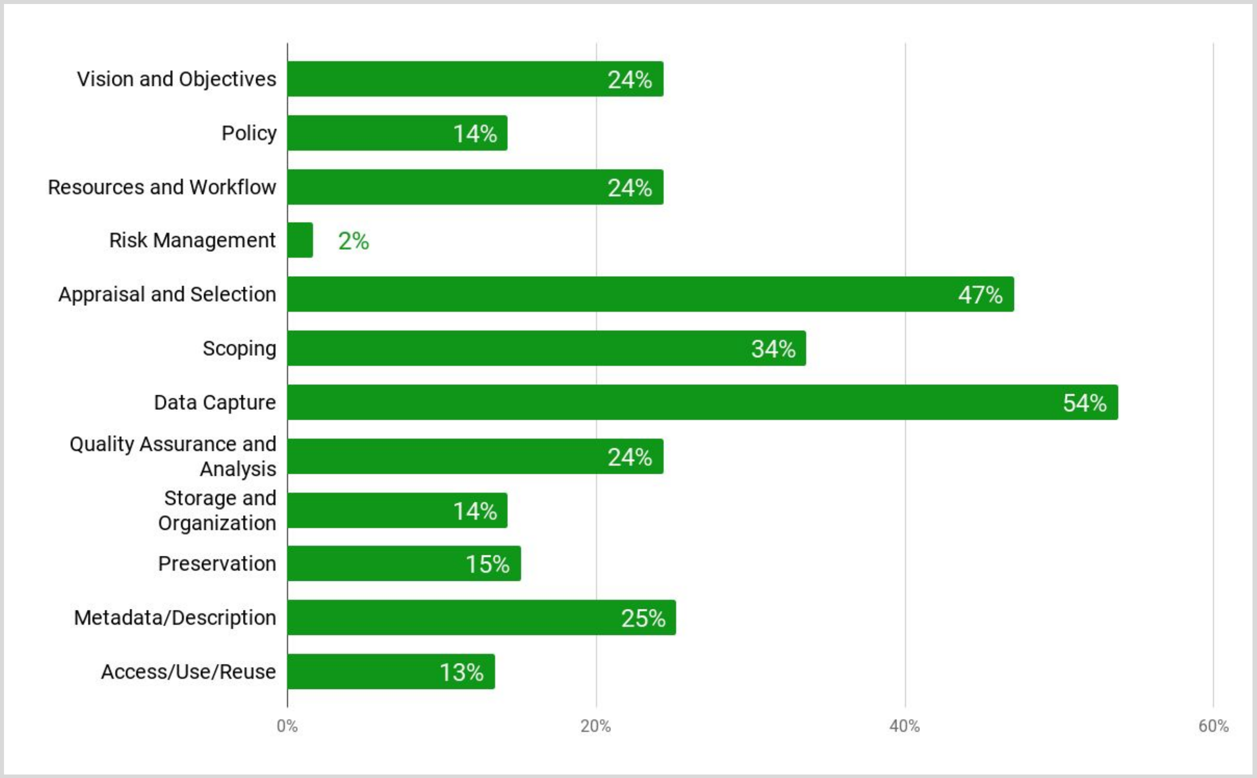

FIGURE 6: Perceptions of most progress in last two years

Team at the Internet Archive, “The Web Archiving Life Cycle Model,” March 2013, accessed August 22, 2018, https://archive-it.org/static/files/archiveit_life_cycle_model.pdf.

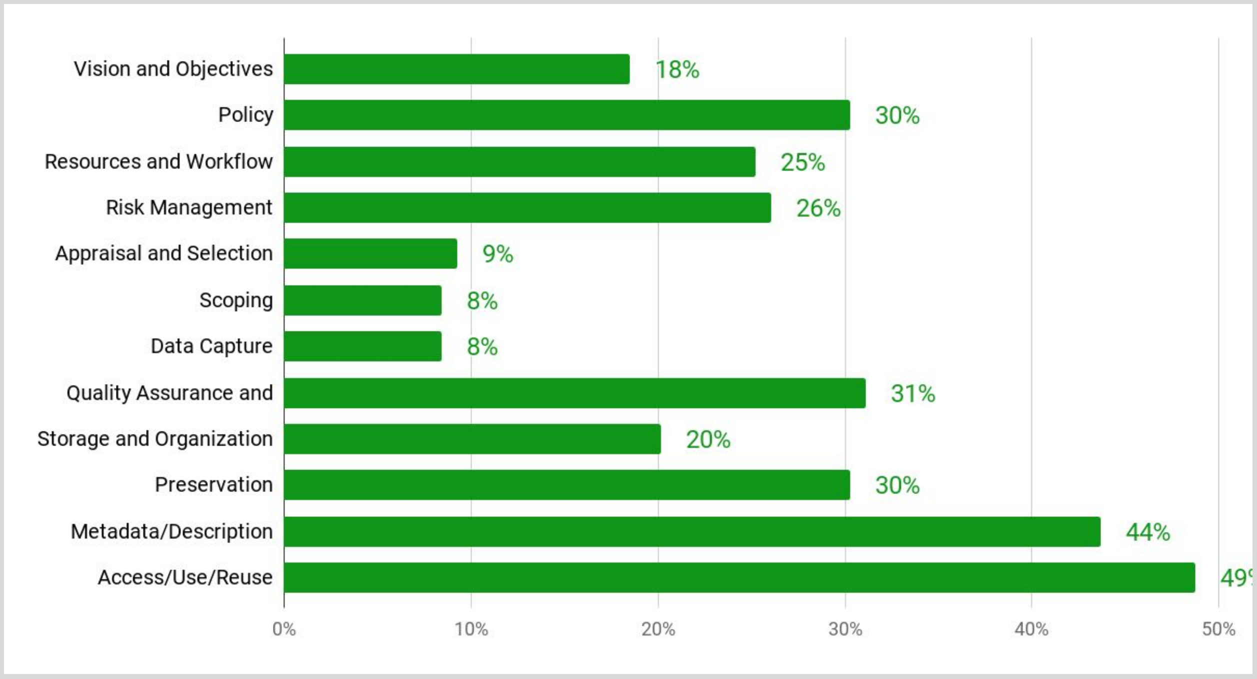

FIGURE 7: Perceptions of least progress in past two years

With respect to dimensions of the Lifecycle Model seeing the least progress, respondents selected Access, Use, and Reuse most often (49%, 58 of 119), Metadata and Description (44%, 52 of 119), and Quality Assurance and Analysis (31%, 37 of 119) most frequently. Compared to the 2016 survey, individual responses and rates underwent slight changes, but the same three dimensions appear as the most frequently selected in the same order. This comparison demonstrates that these three dimensions pose long-running issues to Web archives programs. Particularly around access and use, there is a lack of clarity about how Web archives are to be used post-capture (see Use by Researchers). It will be worth comparing this number to future surveys, particularly now that initiatives such as Archives Unleashed, the IMLS funded program “Continuing Education to Advance Web Archiving,” IIPC’s Training Working Group, and Archive-It Research Services (ARS) exist. Responses also imply that there is some difficulty around using traditional tools for library and archival description for archived Web material.

Staffing, Program Development, and Skills

Respondents were asked to report how much full time equivalent (FTE) staff time is dedicated to their Web archiving programs. Sixty-seven percent of institutions (80 of 119) reported dedicated 0.25 FTE, an increase from 2016 (58%, 49 of 84). A few institutions (9%, 11 of 119) reported 1 FTE to Web archiving activities. Most other options saw small changes, though institutions reporting 0.5 FTE has fallen between 2016 (13%, 11 of 84) and 2017 (7%, 8 of 119). Though this question was not asked directly in 2013, the chart below projects an answer based on free-text answers in that year’s survey.

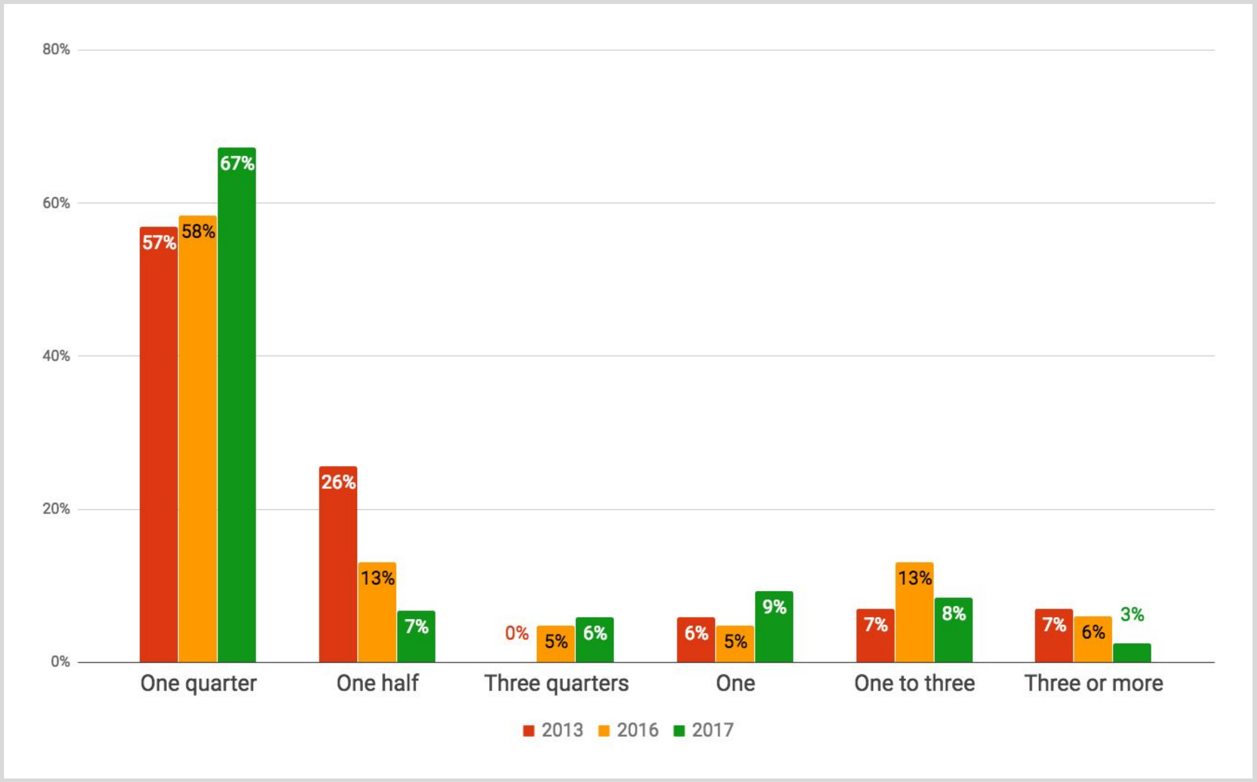

FIGURE 8: Full time staff (FTE) dedicated to Web archiving

Of the 14 institutions who reported being in the piloting or testing phase of their program, 79% (11 of 14) are devoting 0.25 FTE to Web archiving, which is consistent with the 85% (11 of 13) respondents that started their program in 2016 and devote 0.25 FTE. However, even more mature Web archiving programs often only devote the time of 0.25 FTE to their programs. Of the 20 respondents whose programs launched between 2011 and 2015, 75% (15 of 20), as well as the 16 respondents who launched their programs between 2006 and 2010, 56% (9 of 16) were also supported by 0.25 FTE. These responses show that most institutions do not or cannot devote more than 0.25 FTE to their Web archiving programs, programs which take considerable amount of effort to manage. Conversely, it could be the case that mandates at individual institutions are narrow enough that 0.25 FTE is appropriate. It would be worth adding to future iterations of this survey questions about attitudes toward the amount of FTE devoted to Web archiving in order to draw more concrete conclusions.

Question 13 asked “What are the top considerations for the development of your web archiving program? Choose three.” Respondents chose most often “Access and Use (e.g., researcher interactions, web analytics, use cases)” (70%, 83 of 119), “Cost (e.g., budgeting, services allowance utilization, staffing level requirements)” (59%, 70 of 119), and “Quality (e.g., accuracy, completeness, replay fidelity)” (51%, 61 of 119). Of these, Access and Use saw a large increase compared to the 2016 survey (60%, 52 of 87), while the other two categories were consistent with the previous survey’s results. Another notable change in the data was an increase in respondents concerned with loss. Thirty-two percent (28 of 87) of respondents selected “Loss (e.g., link and/or reference rot of archived resources)” as a top consideration in 2016, and that figure increased to 40% (48 of 119) in 2017.[10]

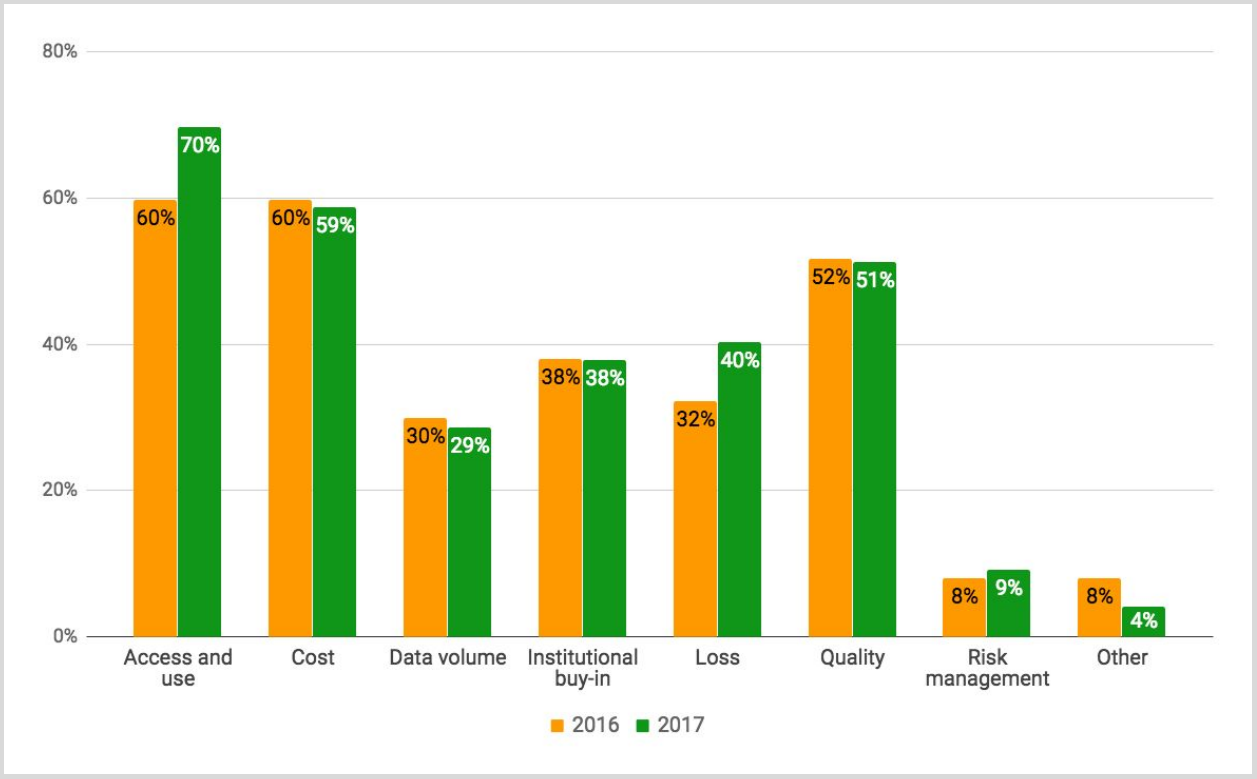

FIGURE 9: Top considerations for the development of respondents’ programs

The survey also asked participants to select the top three staff skills essential to developing a successful Web archiving program. The top three skills chosen by respondents are “Archiving tools (e.g., configuring or operating web archiving tools)” (69%, 82 of 119), “Appraisal and selection (e.g., determining what web content to collect)” (61%, 72 of 119), and “Quality assurance (e.g., analyzing and troubleshooting web archive quality issues)” (49%, 58 of 119). These responses reflect the same trend as in 2016. Interestingly, the only skill that saw more than a 3% change between 2016 and 2017 is “Domain expertise (e.g., knowledge of subjects that are the focus of web archiving),” which saw a 5% jump between the two surveys. These responses indicate that the importance of various staff skills within a Web archiving program has not changed demonstrably from the previous survey.

[10] Perhaps relatedly, professionals across LIS fields and beyond held discussions with respect to loss as it pertains to endangered public data in 2017, with the Internet Archive’s quadrennial effort to preserve websites of the U.S. federal government before a presidential transition receiving particular attention.

Problematic Content Areas

Areas considered most concerning to respondents saw some changes over previous years. Respondents were encouraged to select all that applied, and of those, four categories saw selection rates of over 50% of respondents: Interactive Media (76%, 90 of 119), Social Media (71%, 85 of 119), Video (63%, 75 of 119), and Databases (57%, 68 of 119). Of these, Interactive Media clearly rose in popularity over the other categories, with a 14% increase compared to the 2016 survey, and a 20% increase over the 2013 survey. Other categories saw declines since 2013. Audio Content, for example, held steady between 2016 (33%, 28 of 86) and 2017 (33%, 39 of 119), but overall saw a decrease from 2013 (45%, 39 of 86). Blogs, likewise, only modestly declined from 2016 (21%, 18 of 86) to 2017 (17%, 20 of 119), but the category continued its downward trend from 2013 (36%, 31 of 86). Databases, Video, and Social Media also continue to see downward trends.

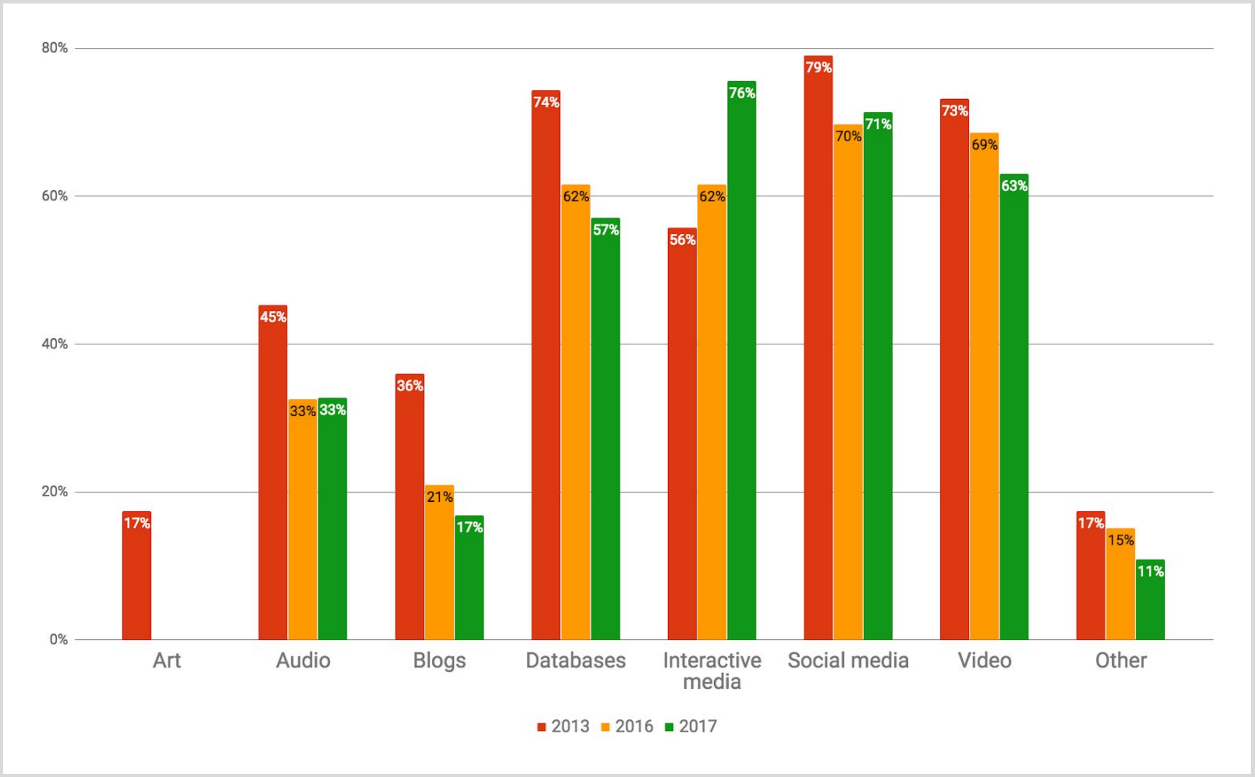

FIGURE 10: Content type provoking concern over capacity to archive

The implications of these trends are difficult to pinpoint. Though Interactive Media is its own category, it should be noted that Social Media and Databases often have some element of interactivity to them. Video, while not interactive, is a complex format, including both time-based audio and video streams and with a Web infrastructure where the largest video hosting sites employ multiple domains to host different parts of what a user would consider one video file.

Attitudes Toward Collaboration

The survey asked respondents about their current collaborative practices with other institutions and their interest levels in such collaborative initiatives. With respect to current collaboration, a small, though slightly increased, proportion answered in the affirmative between 2013 (18%, 16 of 89) and 2017 (23%, 27 of 119). A larger proportion indicated that they did not collaborate in 2017 (43%, 51 of 119), which is significantly lower than in 2013 (82%, 73 of 89). For those respondents who answered “yes,” the survey further asked in what areas collaboration occurs. There was a relatively even distribution between the options provided, though one emerges as a significant leader: eighteen “yes” respondents indicated collaboration around collection development. With the founding of collaborative initiatives like the Ivy Plus Libraries Web Resources Collection Program, the Federal Web Archiving Working Group, and the forthcoming CobWeb, which approach collection development collaboratively by pooling resources and leveraging subject expertise from multiple institutions, future surveys may see a continued rise in collaborative collection development for Web archives.

FIGURE 11: Organizations involved in collaborative collecting

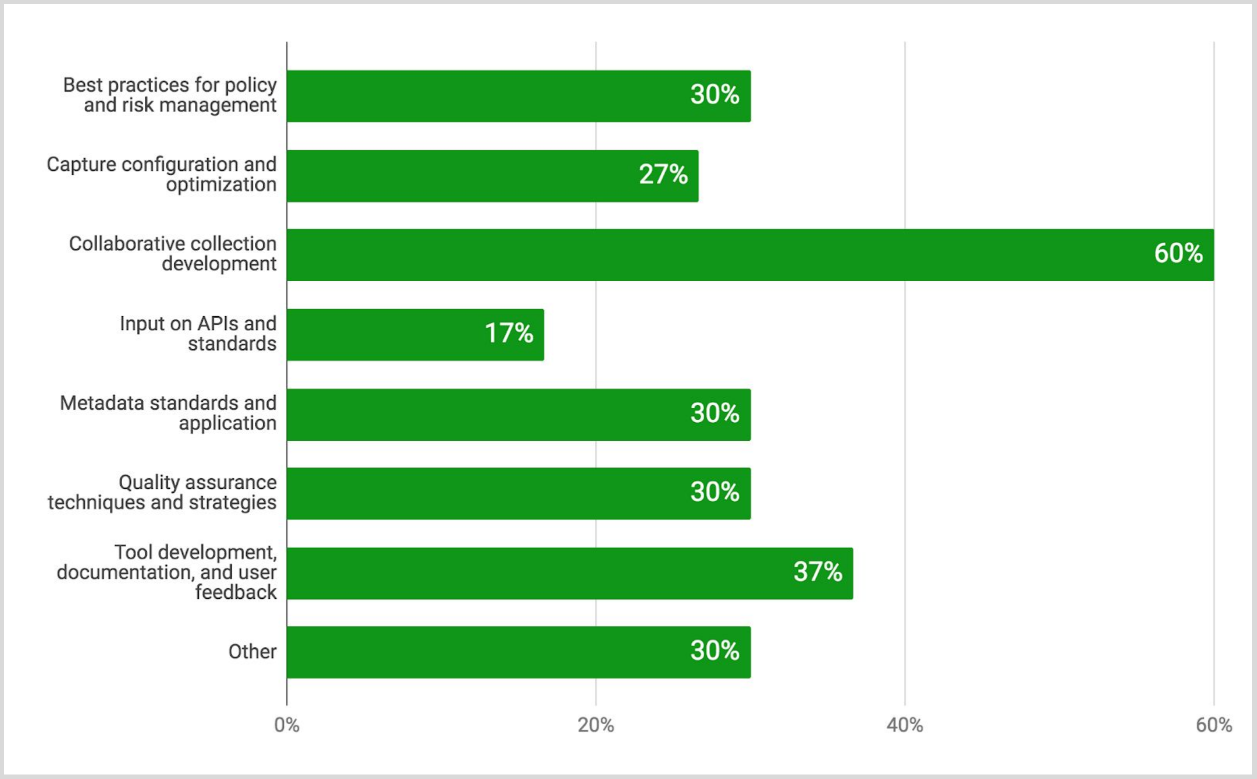

FIGURE 12: Areas in which institutions collaborate

Respondents were also asked to select the barriers they encountered with respect to collaboration. In this, three categories saw modest declines. Institutions pointing to Still Being in Planning or Pilot Stage as a blocker decreased from 30% (25 of 83) in 2016 to 27% (32 of 119) in 2017; Lack of Institutional Support decreased from 20% (17 of 83) in 2016 to 17% (20 of 119) in 2017; and Lack of Time to Spend on Collaborating declined from 75% (62 of 83) in 2016 to 66% (79 of 119) in 2017. The decreased selection of Lack of Time to Spend on Collaborating between 2016 and 2017 is potentially positive if it indeed becomes a trend over additional surveys. Respondents who selected “Other” (20%, 24 of 119) were asked to specify with a free-text field. While free-text responses ranged, 25% (6 of 24) mentioned some variation on lack of staff or staff expertise as a barrier.

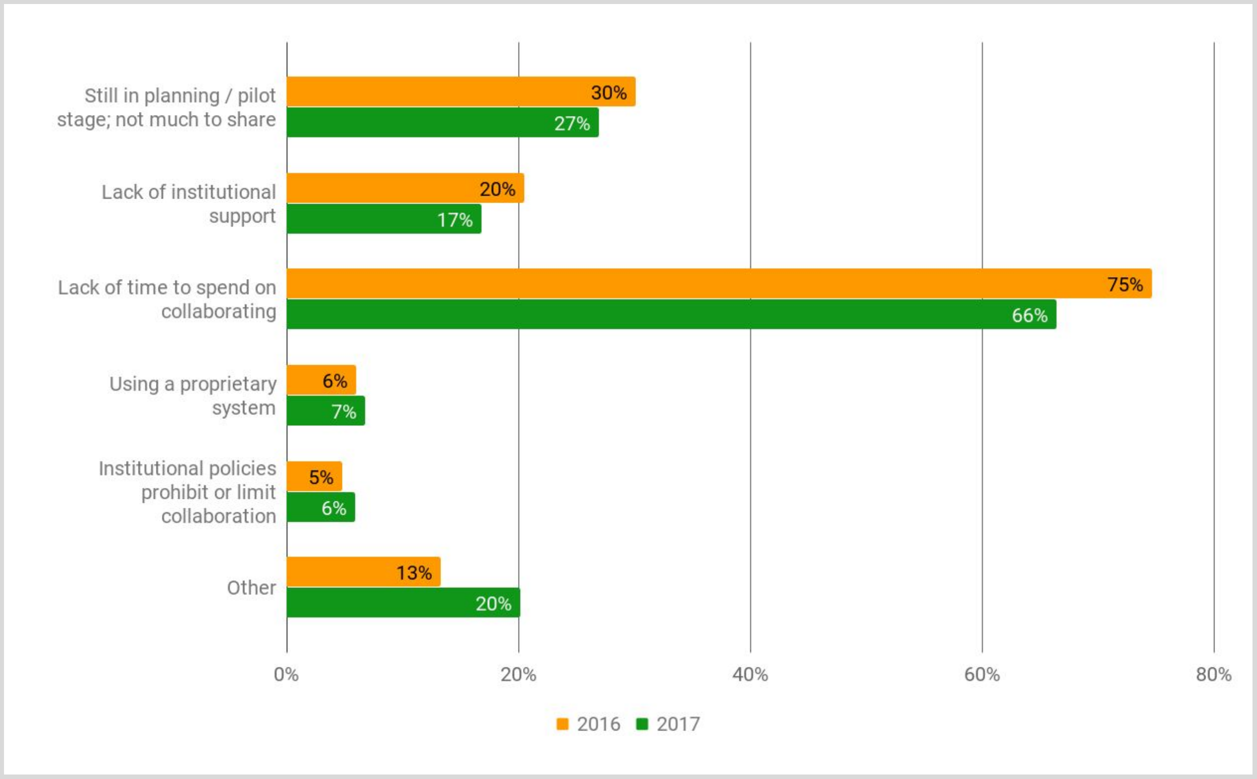

FIGURE 13: Barriers to collaborative collecting

TOOLS AND SERVICES

This section examines the tools and services that institutions use for Web archiving, with particular emphasis on which tools are used locally and with external services, which API tools are used for social media archiving, and whether or not data is transferred.

Local and External

The 2017 results for local and external capture strengthened trends seen in previous years. “Local capture” refers here to Web archiving done in-house; by contrast, “external capture” refers to capture done by an external service provider. These two modes of harvesting may use the same tools to capture data, but may do so in different ways. The vast majority of institutions are still capturing Web content using external services (87%, or 103 of 119), although this is slightly down from 94% (74 of 79) in 2016, compared with 79% (65 of 82) in 2013 and 65% (47 of 63) in 2011. By contrast, 38% (45 of 119) of institutions are capturing locally, up from 36% (29 of 79) in 2016. The percentage of institutions capturing at both local and external levels has remained stable at 30% (36 of 119) in both 2017 and 2016 (24 of 79). This suggests that while external services still dominate the field of Web archiving, institutions have also continued local capture at a stable rate.

The stabilization trend is evident when examining the individual tools that institutions chose for capture. As Figure 14 below shows, 94% (97 of 103) of institutions responding to this question capture content with Archive-It. Archive-It has been the favored external service provider since NDSA began recording these statistics in 2011, when 72% (36 of 50) of responding institutions were using it. Use of Archive-It has increased as more institutions embark upon Web archiving; in 2013, 71% (53 of 75) institutions reported use, and in 2016, 87% (69 of 79) institutions reported use. A scattering of institutions continue to use external services such as Hanzo Archives (2%, or 2 of 103) and OCLC Web Harvester (2%, or 2 of 103). Usage of these services stayed relatively stable as adoption of Archive-It has steadily increased.

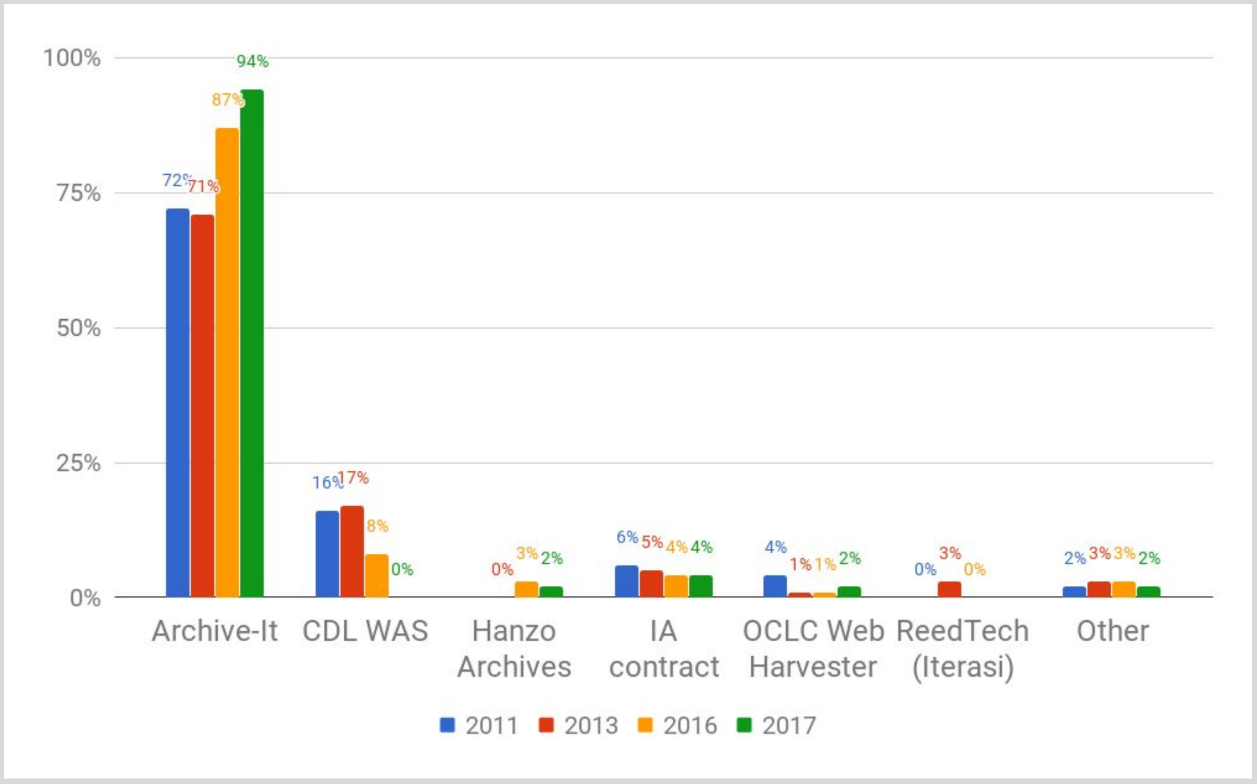

FIGURE 14: Tools used for external capture services

Tools and services used in local capture also bore witness to wider use, albeit to a lesser extent (Figure 15). For instance, the most commonly used tool historically has been Heritrix (one of the tools also used by the external Archive-It service), which has seen a steady increase in usage since the beginning of this survey (24% in 2011, 29% in 2015, 31% in 2016, and 38% in 2017). It is important to note, however, that this data may be inflated due to potential confusion between running a local implementation of Heritrix versus using Archive-It. Although HTTrack has seen similar statistics to Heritrix in the past, its usage dipped this year to 9% (4 of 45) of respondents. Other historically lesser-used tools also saw a dip in use this year, including Wget, down to 13% (6 of 45) from 17% in 2016, and Adobe Web Capture down to 4% (2 of 45) from 10% in 2016, while others ceased to be used entirely (Grab-a-Site, Teleport Pro, WAIL, and Web Curator Tool). By contrast, Webrecorder saw an explosion in use. Because this tool did not exist before the 2016 survey, which saw 21% usage, no data exists for it in earlier NDSA reports. Since 2016, however, it has increased in usage to 51% (23 of 45) of respondents in 2017. Many institutions in 2017 indicated that they use both Webrecorder and Heritrix/Archive-It for capture, indicating that institutions are diversifying their Web archiving toolkit to capture different types of Web material.

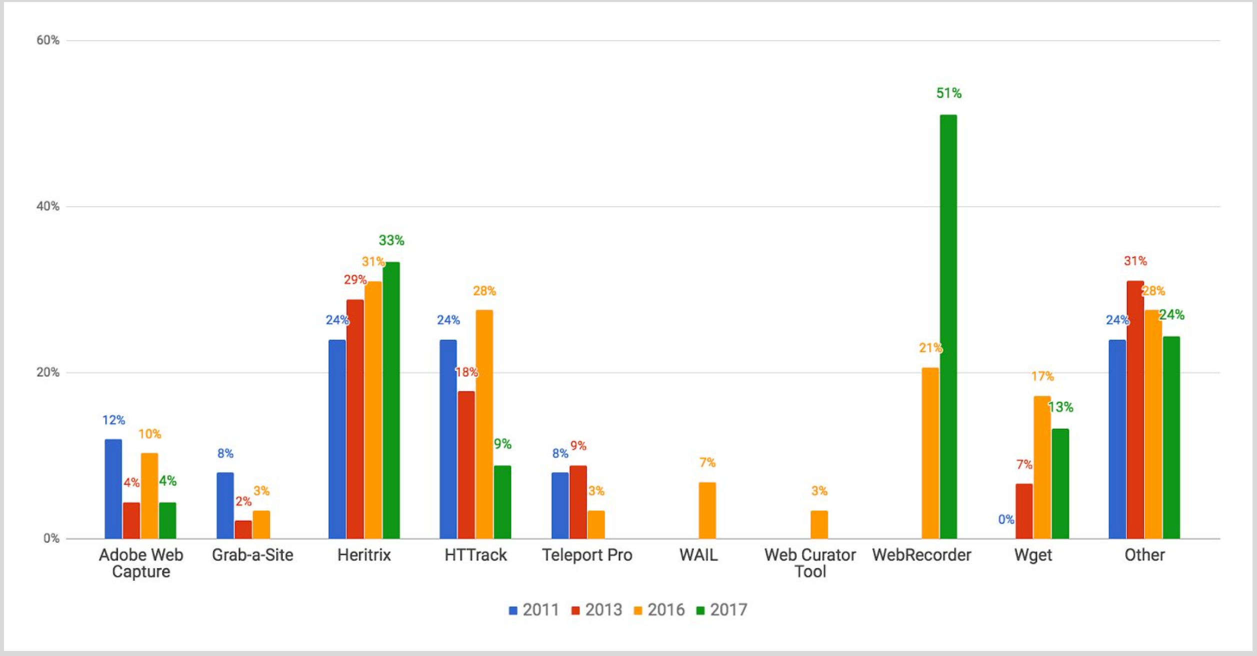

FIGURE 15: Tools used for local capture

Social Media Archiving with API Tools

2017 marks the first year that the NDSA asked about social media archiving using API tools in this Web archiving survey. Although most institutions are still not capturing social media in this manner, 8% (10 of 119) reported that they are.[11] Of these ten respondents, seven use twarc, two use ArchiveSocial, one uses Social Feed Manager, one uses lentil,[12] and one uses Twitter Archiving Google Sheet (TAGS). This suggests that the majority of social media archiving with APIs focuses on Twitter, since both twarc and TAGS harvest tweets. All of the above tools are open source and freely downloadable for local use with the exception of ArchiveSocial, which is a service costing at minimum $199 per month.

[11] These statistics do not include social media archiving using other capture tools such as Heritrix or Webrecorder as these are not API based tools.

[12] After these results were gathered in 2017, Instagram released a new API that no longer supports lentil functionality. As a result, lentil is no longer supported by its developers. "README.md," NCSU-Libraries, February 9, 2018, accessed August 22, 2018, https://github.com/NCSU-Libraries/lentil/blob/master/README.md

Data Transfer

Institutions may choose to transfer their Web data from an external service to a locally- or remotely-hosted repository for a variety of reasons, such as ease of access or preservation redundancy. Like previous surveys, the percentage of institutions who choose to replicate their data hovered around 19% in 2017, despite the increased use of external tools (Figure 16). Those who did replicate did so increasingly at the local level, at 80% (16 of 20), up from 59% in 2016.

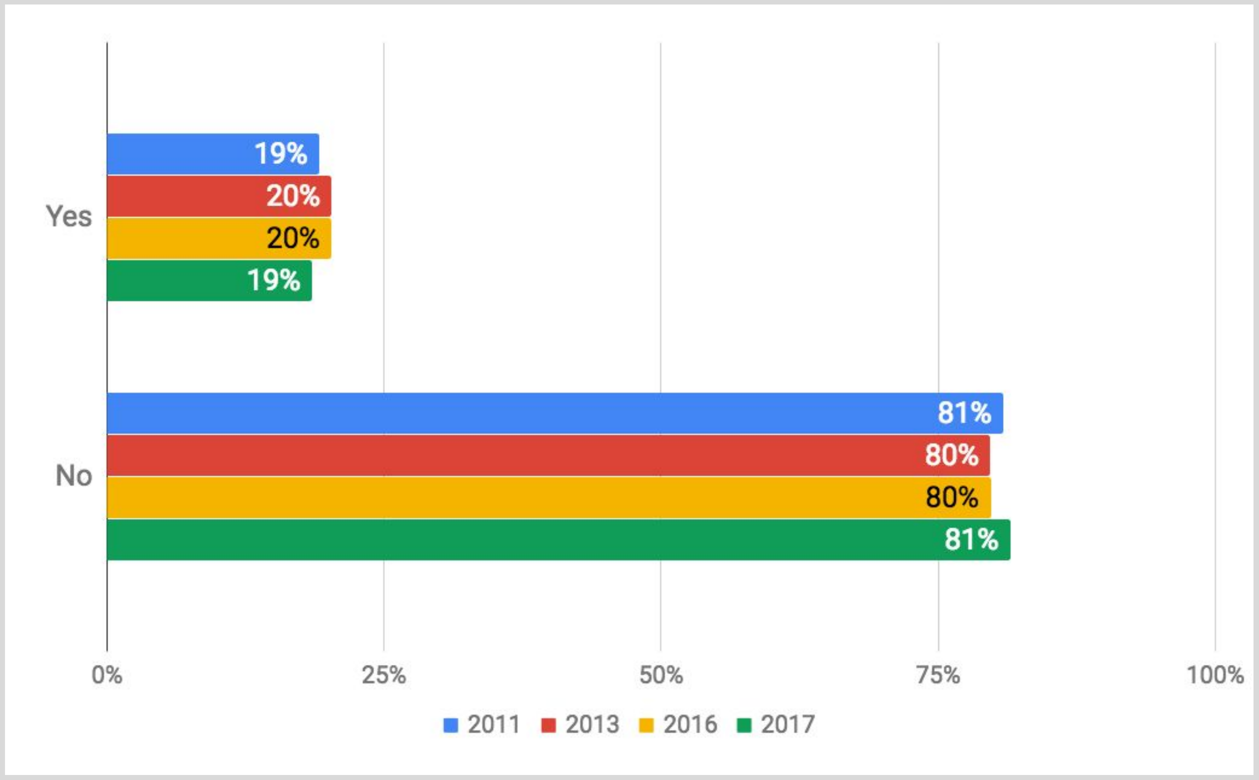

FIGURE 16: Organizations that choose to replicate data

When asked why they had not replicated their data to another location, the percentages for the provided options of "trust data capture service provider," "building local infrastructure," “no place to store/maintain it," and "not sure what we'd do with it once we got it" all were lower than in 2016. The only category that increased was “other” (Figure 17). The reasons respondents provided under “other” for not transferring data from an external services included included lack of time, lack of planning, lack of funding, no staff to manage or work on the activities, and that Web archiving data was not an institutional priority.

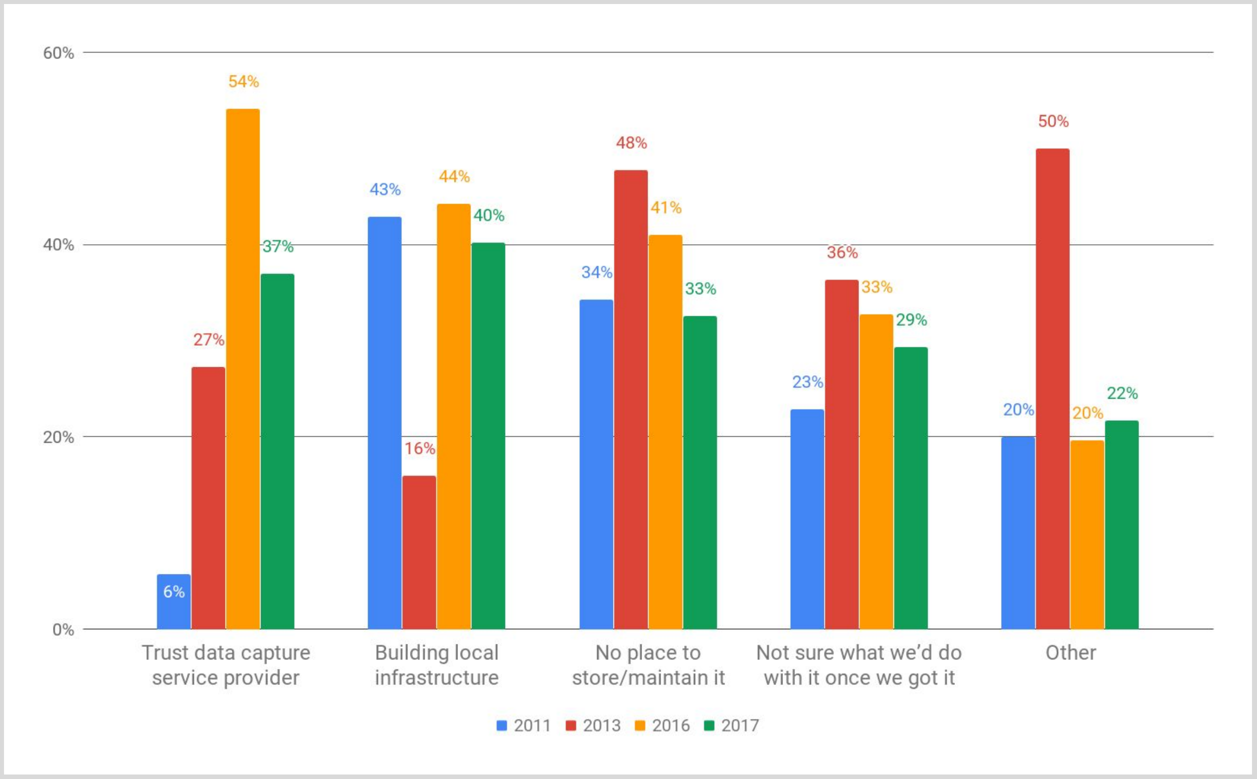

FIGURE 17: Reasons for not transferring data from an external service

Use by Researchers

For the second time, the survey included a question about researchers actively utilizing Web archives to learn more about engagement with users of Web archives. The options were simply: "yes," "no," or "I don't know." As the chart below shows, the responses remained consistent with 2016 findings. "I don't know" is still the most prevalent answer, but this year, less than half, 49% (57 of 117) of respondents, chose this answer as compared to 51% (41 of 80) of respondents in 2016. The number of organizations confident that active researchers are not utilizing the archive climbed slightly from 30% (24 of 80) of respondents in 2016 to 33% (39 of 117) of respondents in 2017.

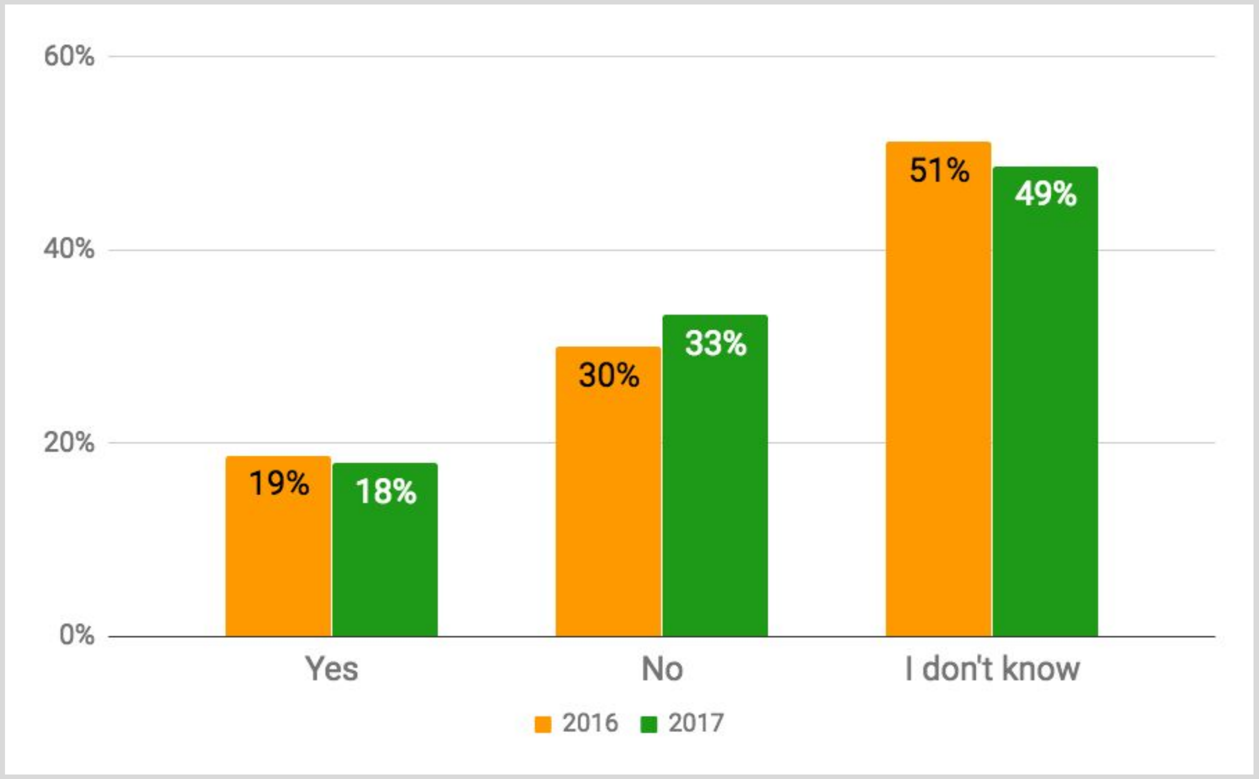

FIGURE 18: Active usage of Web archives by researchers 2016-2017

When asked to explain how it was known that researchers used the archives at these organizations, narrative answers included usage tracking tools such as Google Analytics and first-hand knowledge of students using the Web archives for class projects and academic research projects.

ARCHIVING POLICIES

This section examines institutional policies for Web archiving programs, especially those that dictate notification and permission requirements, access embargos, robots.txt files, copyright, and social media archiving.

Notification and Permission

Results from the 2017 survey show an increase in this consensus towards permission policies (Figures 20-22). Seventy percent (71 of 101) of institutions capturing content do not seek permission or attempt to notify the content owner that their website is being archived. This number is up from 67% (46 of 69) in 2016, and 58% (42 of 73) in 2013. This approach remains relatively consistent regardless of whether the data they are capturing will have restricted access or public access. In fact, 75% (76 of 101) of respondents do not seek permission to capture websites when access will be restricted, and 71% (72 of 101) do not seek permission to capture sites when access will be public. This suggests that institutions with a Web archiving program have become more comfortable archiving without permission and notification in the past six years. As the 2016 survey and Figure 4 suggest, this may also indicate that institutions are increasingly archiving their own content, or content for which they do not feel the need to request permission.

FIGURE 19: Notify/permission approach: capturing 2011-2017

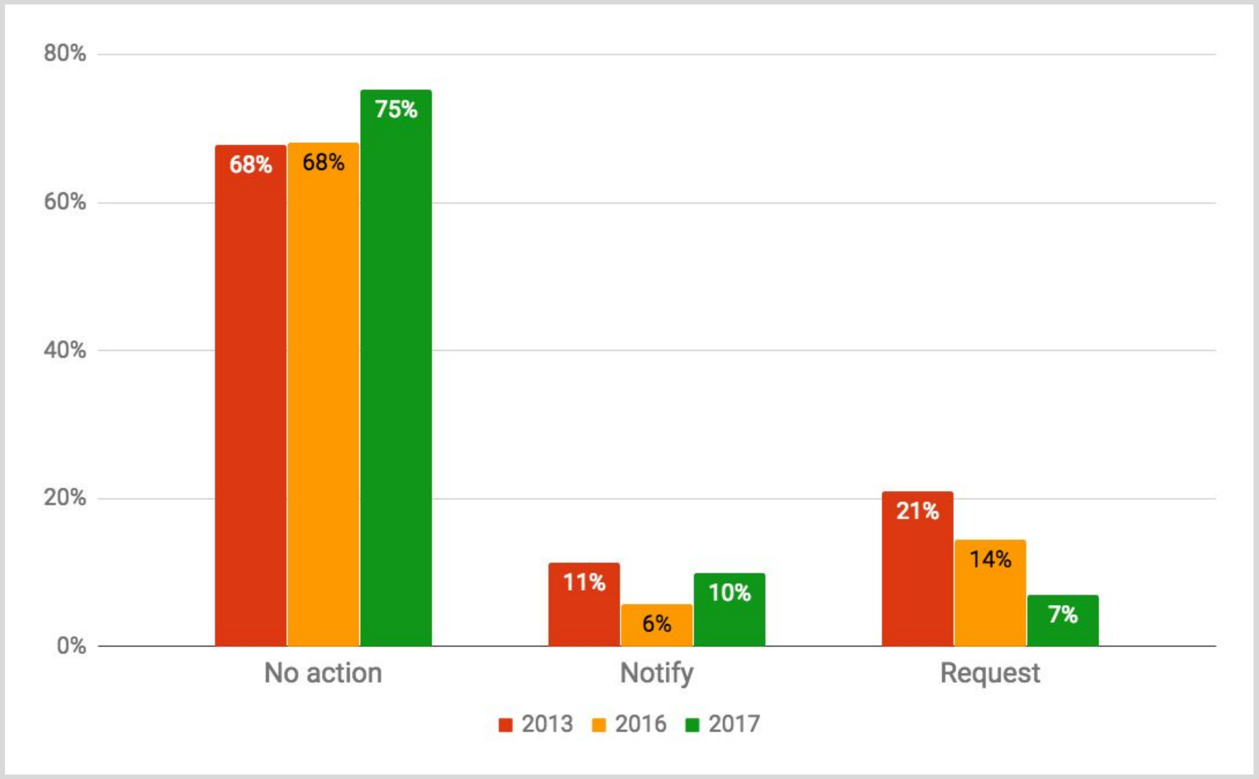

FIGURE 20: Notify/permission approach: restricted access 2011-2017

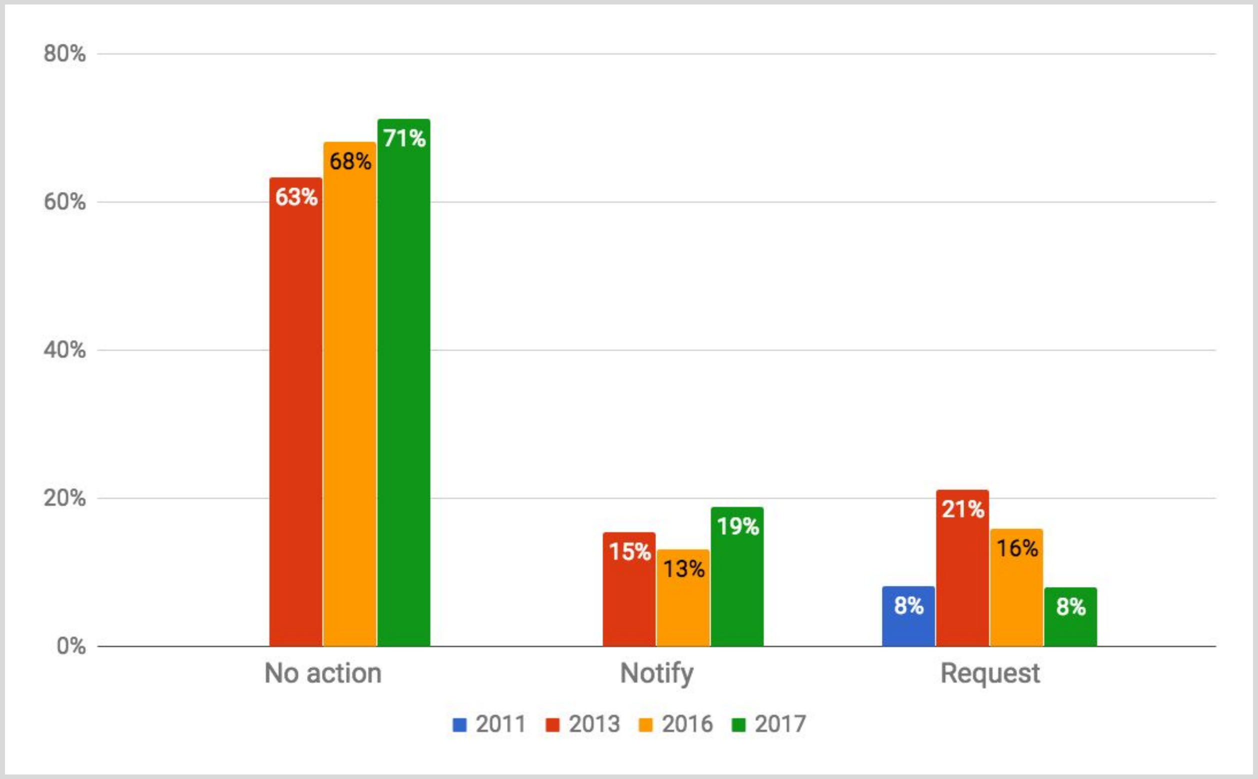

FIGURE 21: Notify/permission approach: public access 2011-2017

Additionally, 91% of respondents (106 of 117) reported that they had never received a request to take down or stop crawling content for which they had not received explicit permission (Figure 23). Although this is a new question for the NDSA survey with no comparable historical data, it does support the hypothesis that institutions are more comfortable crawling sites without permission or notification since only 9% (11 of 117) have ever received a takedown request.

FIGURE 22: Request for take down or stop crawling content 2017

Access Embargo

Institutions may choose to put an embargo on archived Web content, which restricts public access to archived sites for a set amount of time. This may be done for many reasons; for instance, embargoing an archived site might reduce competition and/or confusion with the live site, or protect the privacy of the creator or subjects of the site.

The percentage of institutions that reported using embargos remained similar to those from 2016 (Figure 24). The distribution of embargo usage remained steady: 14% (16 of 116) use embargos, 66% (77 of 116) do not use embargos, and 14% (16 of 116) are considering using embargos, while 6% (7 of 116) are dark archives.

Institutions reported significant variation between their own policies dictating embargo lengths, although the sample size for this question was so small (11 respondents in 2016, 18 in 2017) that responses had a significant impact on percentages. In 2017, the question was also rewritten to include specific options for embargos that lasted over a year. In 2016, three institutions (27%) reported having policies that outlined embargo periods that lasted less than six months, whereas in 2017, no institutions reported using an embargo that had such a short duration. Instead, most respondents in 2017 implemented embargo periods between six months and a year (39%, 7 of 18). In 2016, 55% (6 of 11) reported using “other” embargos, including embargos for over one year. In 2017, options were given for one to two years (17%, 3 of 18) and two or more years (6%, 1 of 18), as well as “other” (39%, 7 of 18). Responses for “other” expanded upon varying embargo policies for different content such as dissertations and content marked for internal use. Several respondents also indicated that they lacked a set embargo policy, but were in the process of developing one.

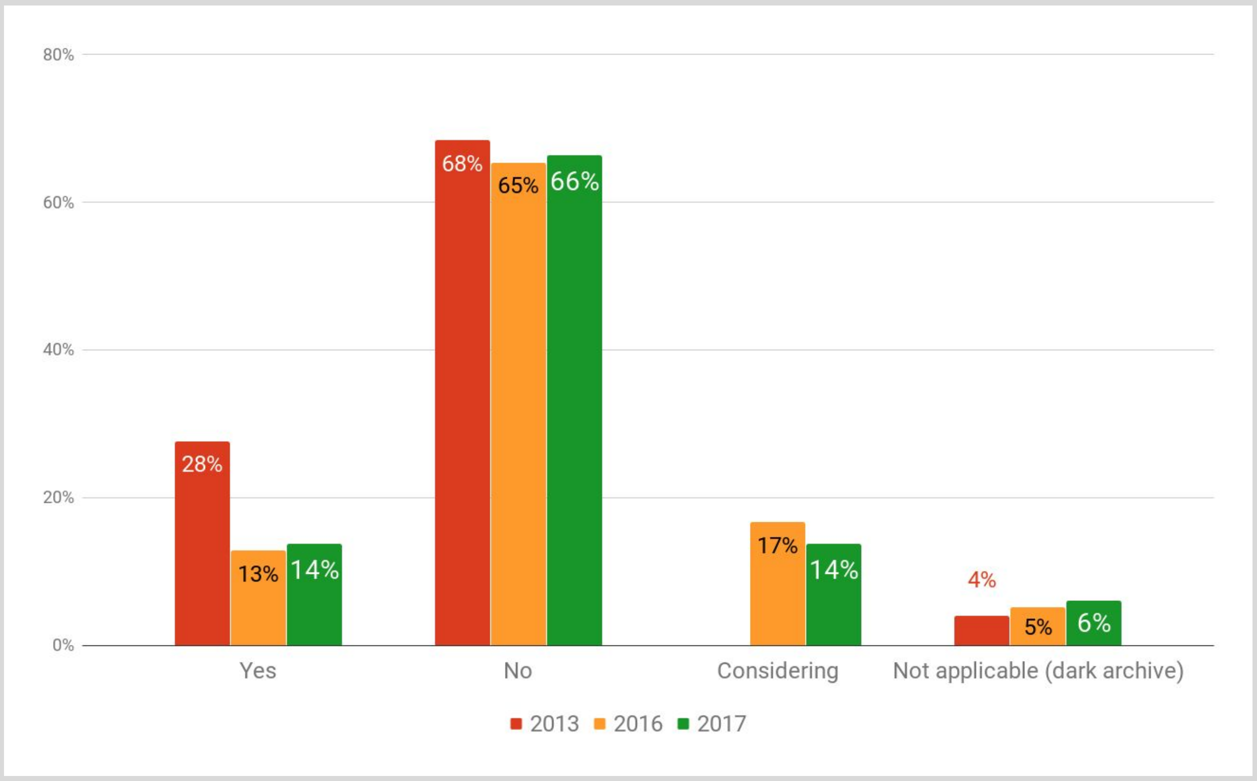

FIGURE 23: Use of embargos in organizational policies 2013-2017

Robots.txt Policies

Robots.txt files are associated with websites to restrict crawling via a machine-readable mechanism. Usually, these files tell crawlers which portions of the site they can or cannot crawl. This impacts Web archiving campaigns because Web archivists choose whether or not to ignore robots.txt files when they crawl sites for content.

As with recent NDSA Web Archiving surveys, the majority of institutions in 2017 respect robots.txt files on a conditional basis. Sixty-four percent (74 of 116) stated that they “sometimes” respect these files. These responses decreased slightly from 2016, when 76% (60 of 79) of institutions reported using this conditional approach. In 2017, more institutions took an either-or stance on robots.txt files: 11% (13 of 116) reported to “always” respect them, while 10% (12 of 116) reported to “never” respect them. Additionally, 17 institutions (15%) reported they “didn’t know” their policy toward robots.txt files, a significant increase from last year’s 0%. These numbers may suggest that institutions have developed a less customized approach per each site they crawl as the scope of their Web archiving increases. It is impossible to definitively conclude this without a more detailed survey of robots.txt policies, especially given the very small data pool of responses.

When asked why they "sometimes" respected robots.txt files, most institutions again stated that they had special access or copyright privileges to the content (65%, compared to 61% in both 2016 and 2013). An overwhelming majority (64%) also stated they adopted this policy in order to capture "essential" content, a steady increase from 52% in 2016 and 43% in 2013. By contrast, fewer institutions reported having secured permission before ignoring robots.txt files (40%, 29 of 72 in 2017; down from 48% in 2016 and 50% in 2013). Respondents also indicated in their write-in responses that they ignored robots.txt files when the robots files were likely created automatically by the CMS used to build the site rather than by the site creators.

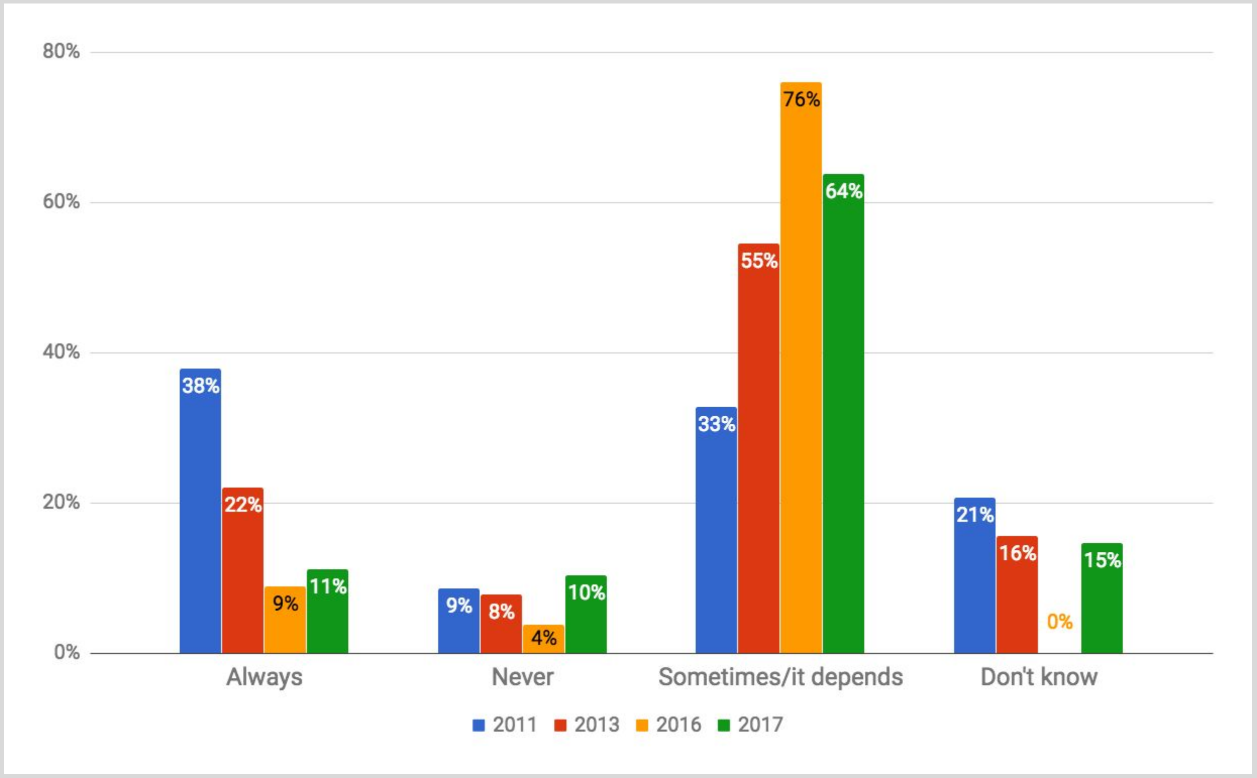

FIGURE 24: Policies for respecting robots.txt

Copyright and Policy Development Resources

As in previous surveys, the most commonly used resources for Web archiving policy development are existing policies of similar organizations. Sixty percent of institutions report consulting others' policies while developing their own (54 of 90), a number that has remained relatively stable over the history of this survey. Institutions also consult the ARL

Code of Best Practices[13] (29%, 26 of 90, down from 40% in 2016), as well as legal counsel (31%, 28 of 90, down from 40% in 2016). Others also consult Section 108 Study Group[14] (17%, 15 of 90), statutory authority (14%, 13 of 90), and previous NDSA surveys (20%, 18 of 90).

Social Media

Social media archiving policies were slightly more common in 2017 than they were in 2016, with 11% (13 of 116) of institutions now reporting policies in comparison to 7% (6 of 82). This number is still not as high as it was in 2013—the first year the NDSA Web Archiving Survey asked this question—when 50% more institutions reported adoption of social media policies (24%, 19 of 78). As the 2016 survey team hypothesized, this may be due to integration of social media with general Web archiving and a potential lack of distinction between policies. However, it may also be due to a surge of interest in social media archiving in 2013 coupled with the subsequent realization of increasing technological challenges associated with this process. Regardless of the reason in the plunge of social media archiving policies, it is encouraging to see that they are on the rise again.

[13] The Association of Research Libraries (ARL), "Code Of Best Practices In Fair Use For Academic And Research Libraries," January 2012, accessed August 22, 2018, http://www.arl.org/storage/documents/publications/code-of-best-practices-fair-use.pdf.

[14] "The Section 108 Study Group Report," March 2008, Accessed August 22, 2018, http://www.section108.gov/docs/Sec108StudyGroupReport.pdf.

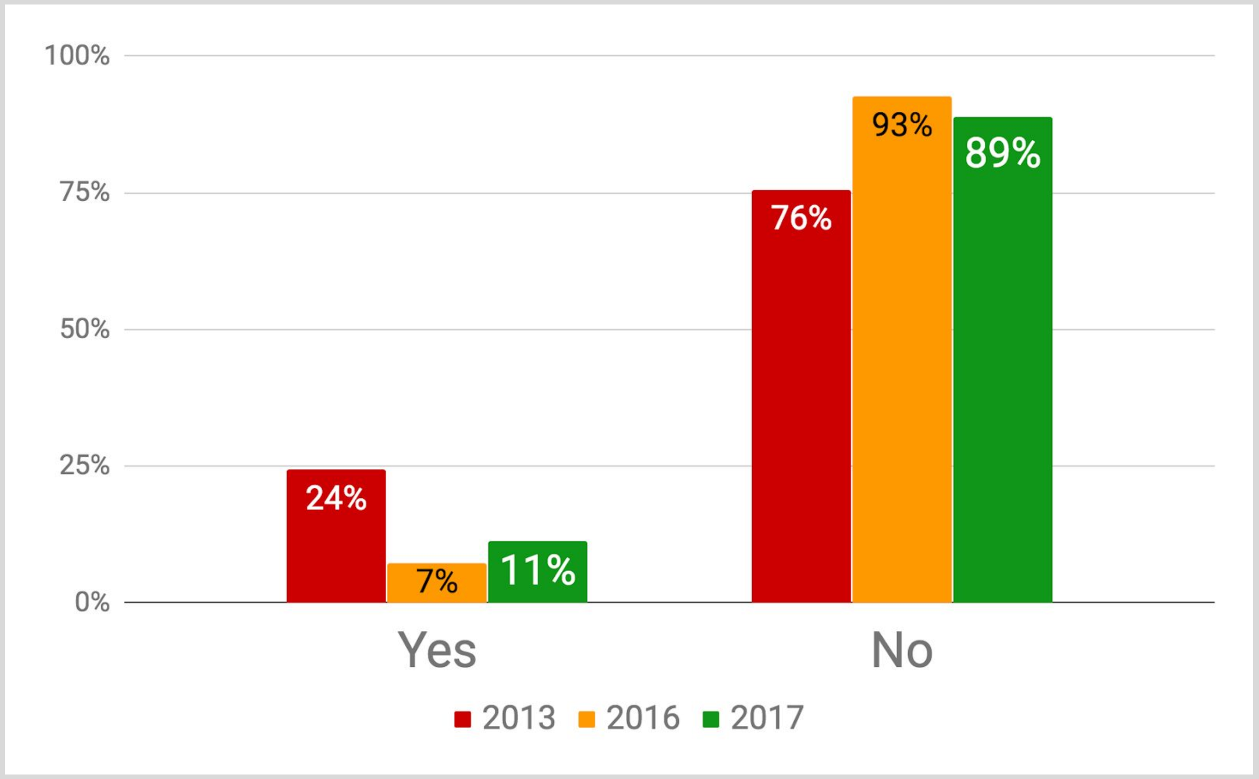

FIGURE 25: Adoption of social media policies 2013-2017

CONCLUDING OBSERVATIONS

The state of Web archiving as represented in the 2017 survey illustrates significant growth in key areas, with continuing stagnation in others. Diversification of the field, maturation of programs, and technological developments presented areas of progress for the profession while access to archived content and institutional support for program expansion remained relatively unchanged from prior survey years.

Organizational Representation

Throughout its history, the NDSA's Web Archiving Survey has sought to track the development of the profession. This year's survey highlighted several areas in which the Web archiving field has evolved.

While the majority of Web archiving continues to occur in the college/university setting, a significant number of public libraries have joined the community. This welcomed expansion of organizations involved in the practice is due to the aforementioned Community Webs project. Continued diversification of the field will serve to expand the communities documented in collections of archived Web content.

Technological Diversification

The majority of responding organizations reported utilizing multiple tools to capture content successfully. The diversification of tools utilized is exemplified by the development and adoption of the Web capture service, Webrecorder. A notable 51% (23 of 45) of organizations reported using Webrecorder, which was publicly launched in 2016 as a browser-based tool to allow for the capture of content difficult to capture via traditional link-based crawling. While some organizations are likely ingesting their Webrecorder WARCs into their overall Web archive, or ingesting this data into other repositories or services, future surveys could aim to better understand how institutions are providing access to content captured via Webrecorder. Archive-It continued to be the preferred external service for harvesting Web content, with 94% (97 of 103) of respondents using this service. While the vast majority of respondents are utilizing Archive-It, few (20 of 108) organizations reported downloading their WARCs for local preservation or access, continuing a trend denoted in previous surveys.

Advances in Description

More organizations are providing catalog records for archived Web content both at the Web collection level and item level. While less than half of respondents describe Web archives via catalog records (36%, 36 of 99 for collection records; 23%, 23 of 99 for item records), the proportion has increased since 2013. On the other hand, 44% (52 of 119) of respondents chose Metadata and Description as a dimension of the Lifecycle Model for which they had made the least progress. After the conclusion of the 2017 survey, OCLC Research released a set of publications titled “Descriptive Metadata for Web Archiving,” one of which recommends which data elements to use in describing Web archives and crosswalks to popular metadata standards, including Dublin Core, EAD, and MARC 21.[15] If future surveys see a continued rise in the use of catalog records to describe Web archives, it may be useful to include questions about respondents’ use or awareness of these recommendations.

Policy

This year’s survey marked the first time respondents were asked if they had ever received a request to take down or stop crawling content for which they had not received explicit permission. Notably, only 9% (11 of 117) reported ever being asked to remove archived content. This heartening statistic will hopefully encourage some existing programs to consider expanding the scope of their collecting to include content created outside of their institution or to implement more flexible policies around permission and access.

[15] Dooley, Jackie and Kate Bowers, “Descriptive Metadata for Web Archiving: Recommendations of the OCLC Research Library Partnership Web Archiving Metadata Working Group.” February 2018, accessed August 22, 2018, https://www.oclc.org/content/dam/research/publications/2018/oclcresearch-wam-recommendations.pdf.

ACKNOWLEDGEMENTS

We would like to thank all the participants in the 2017 Web Archiving Survey for their time and willingness to share information about their programs with the broader Web archiving community. Many thanks to NDSA, CLIR, and others for their support and feedback throughout the project, especially the NDSA Coordinating Committee, Jefferson Bailey, Jesse Johnston, Carol Kussmann, Katherine Kim, Christie Moffatt, and Nicholas Taylor for their editorial assistance.

APPENDIX A: Survey Questions

The PDF of the 2017 Web archiving survey questions will be available at http://ndsa.org/publications.